Thesis title:

Reliable and functionally safe AI hardware

Institution:

Sorbonne Université
French National Centre for Scientific Research (CNRS)
Computer Science Laboratory of Sorbonne University's Faculty of Science and Engineering (LIP6)
Sorbonne Center for Artificial Intelligence (SCAI)

Location:

Paris, France

When:

Starting October 2019

Funding:

3 year PhD grant, ~1700€ monthly gross salary, social security benefits included.

Advisor:

Haralampos-G. Stratigopoulos

Context:

Typically artificial intelligence (AI) algorithms run in software on general-purpose central processing units (CPUs) or on giant servers in the cloud using clusters of CPUs. Due to concerns of latency, network bandwidth, and privacy, there now exists a trend to push AI from cloud to edge, where computation is largely or completely performed on distributed Internet-of-Things (IoT) devices. However, a CPU is too large to fit inside an IoT device and it needs far more power than a device battery can provide. In addition, several AI applications require a real-time response. An example is autonomous driving technology where AI algorithms will become a standard especially in infotainment human-machine interfaces (i.e. virtual assistance, natural language interfaces, etc.) and Advanced Driver Assistance Systems (ADAS) (i.e. camera-based machine vision systems, radar-based detection units, etc.). In such applications, performing inference or on-line learning with a neural network running in software on a CPU will result in unacceptable latency. Similar area, power, and latency limitations are observed when graphic processing units (GPUs) and field-programmable gate arrays (FPGAs) are used instead. These limitations, along with other more technical challenges, such as the von Neumann bottleneck and the approaching end of Moore's law, have made it crucial to develop dedicated and customized AI hardware architectures.

The answer comes with neuromorphic computing, a term referring to special purpose very large-scale integration (VLSI) artificial neural network (ANN) implementations that resemble -or are inspired from- biology. The work that followed has led to the advent of VLSI implementation of ANNs that can work as customized hardware accelerators or can offer a much smaller form factor and better energy efficiency, such that they can be used in resource-constrained IoT nodes for near-sensor computation and near-sensor intelligence.

Today, there exist hardware ANN architectures that can be digital, mixed analog-digital, purely analog, or based on emerging non-volatile memories. There exist other categorizations of hardware ANN, for example based on the type (i.e., level-based or spiking), the topology (i.e., feed-forward, radial-basis, convolutional, etc.), and the underlying technology used (i.e., CMOS, CMOS-compatible, post-CMOS).

With the foreseen industrialization and high-volume production of hardware ANNs in the coming years, designing reliable, self-testable, fault-tolerant, and functionally safe hardware ANNs is an emerging topic that is largely unexplored. Hardware ANNs, as any other integrated circuit (IC), are subject to errors occurring during the several manufacturing steps (i.e. process variations and defects). Even if they pass the post-manufacturing step, they may fail later in the field of operation due to ageing and wear-and-tear provoked by environmental stress, such as heat, humidity, and vibration. There is a general belief that errors can be tolerated due to the modularity and high parallelism of ANNs. However, this is not true since most applications will demand compact hardware ANN and errors that render even a small part of the hardware unusable may seriously deteriorate the learning capacity. For example, in Google's Tensor Processing Unit (TPU), which is a systolic-based array design composed of a grid of 256x256 Multiply-And-Accumulate (MAC) units, it is shown that if 0,005% of MAC units is faulty, then the classification accuracy on the TIMID benchmark drops from 74.13% to 39.69% .

In this thesis, we envision developing built-in self-test (BIST) methodologies that equip the AI hardware with the capability to identify and neutralize faulty synapses and neurons. This in turn will enable a faster and more robust learning, and is an important step towards guaranteeing the highly demanded reliability, active fault tolerance, and functional safety requirements. We envision developing versatile BIST circuitry that looks solely at the hardware independently of the cognitive task and data being processed. We will develop and validate by simulation BIST methodologies for ANN architectures at transistor-level. We will also design, fabricate, and demonstrate a hardware ANN with BIST capabilities.

Short Bibliography:

[1] C. Torres-Huitzil and B. Girau, "Fault and Error Tolerance in Neural Networks: A review," *IEEE Access*, vol. 5, pp. 17322 – 17341, 2017

[2] J. J. Zhang,T. Gu, K. Basu and S. Garg, "Analyzing and mitigating the impact of permanent faults on a systolic array based neural network accelerator," in Proc. *IEEE VLSI Test Symposium*, 2018.

[3] L. Anghel, G. Di Natale, B. Miramond, E. I. Vatajelu, and E. Vianello, "Neuromorphic computing - from robust hardware architectures to testing strategies," in Proc*. IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC)*, 2018, pp. 176–179.

[4] S. Yu, "Neuro-inspired computing with emerging nonvolatile memories," *Proceedings of the IEEE*, vol. 106, no. 2, pp. 260–285, 2018.

[5] S. A. El-Sayed, L. A. Camunas-Mesa, B. Linares-Barranco, and H.-G. Stratigopoulos, "Self-Testing Analog Spiking Neuron Circuit, " in Proc. *International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design*, 2019.

Expected skills:

We seek a highly motivated talent with a M.Sc. degree or equivalent in Electrical Engineering or Computer Engineering and with background knowledge on circuit design, computer-aided design tools (e.g. Cadence, Synopsis, Mentor), and technical computing languages (e.g. MATLAB). Knowledge on AI algorithms and applications is a plus.

How to apply:

Send by e-mail a detailed CV to Haralampos-G. Stratigopoulos (e-mail: haralampos.stratigopoulos@lip6.fr).